

BE -09

THE RELATIONSHIP ANALYSIS BETWEEN DISCRIMINANCY POWER AND VALIDITY OF ITEM TESTS

Muhyiatul Fadilah

*Biology Education
State University of Padang*

Abstract

Test is one common instrument, which used by most of college students who focus on experimental study and also most teachers. Items test is arranged refer to learning goal then its is tried out in an examination simulation for parrallel leaners. Based on learners' answer, can be determined the level of item tes quality. There are four description that must be contained by each item test, are proper difficulty index, empirical validity, discriminancy power, and instruments reability. It was found that tester got difficulty to choose one best quality item which contain these description, because there is no sufficient information about criteria for a best quality item test. Eventhough, applying bad quality item cause a bias for learning result measured. Generally, this research aimed to provide an correlation facts among some of these four item criteria, especially between discriminancy power and validity. The population is test instrument for students at senior high school. Twelves samples of test intrument have been chosen using purposive sampling. Data analysed by using correlation technique of SPSS analysis, also descriptively. The result of this study is dicriminancy power correlates toward item validity, with determination coofecient start from 54.8% to 82.6%. Thus, each item description is able to reflect another. Item which has higher discriminancy power value also has higher validity item value.

Key words: Correlation analysis, Discriminancy power, Item validity, Test

INTRODUCTION

Test is a common intrument used by most teachers in assessing students achievement. Multiple choice is one of favourite item form used by teachers because easy to be constructed because it can represents almost learning material which will be tested, easy in scoring, and inform high objectivity. Preparation of test is started with setting the goal, making test blue-print, writing question, review the item tests, assembling become intrument, testing out and determining the item test criteria, are validity, reliability, difficulty index, and discriminancy power.

In fact, most of the tests generated by teachers do not through established procedures. Based on a survey and interview with some of the high school teachers in November 2010, informed that most of teachers create the tests which items are copied from some items bank available without considering items-learning goal suitability, conformity cognitive level, and analyze the item qualitatively or quantitatively.

Qualitative analysis includes considerations of content and construct validity criteria.

Logically review is one way to acquire these criteria. Quantitative analysis is an analysis for assessing test instruments through interpretations of amounts of data with its statistical character (Anastasi and Urbina, 1997, p:172) or an procedure in judgemental upgrading and a procedure for empirical improvement procedure (Popham, 1995, page:195), include items test validity, reliability, difficulty index, and discriminancy power. Of the two types of analysis, this study focuses on items test analysis.

Theoretically, the item analysis is an absolute requirement that must be met by a teachers before items test tested for testee who became the subject of assessment. Facts on the field indicated that most teachers ignore the feasibility of this use. Among the reasons are, 1) the teachers feel burdened by the analysis process so tend to be disregarded, 2) the teacher do not aware for bias of result of assessment generated by the inappropriate items test, 3) teachers feel very confident with the quality of the instruments made so feel no need to conduct further review. In fact, bad quality items test cause faulty interpretation of learning outcomes.

Teachers objections to items analysis also caused any difficulties to determine which item are ready-used, need revision, or being disposed based on its standard criteria assembled from whole of aspects of analysis, are validity, reliability, difficulty index and discriminancy power. Basically, teachers made a success to perform quantitative item analysis, however, slightly information available to judge item quality and status, whether these ready to apply, need revision, or canceled.

Test arrangement is also one of common activity carried out by college students during Learning Assessment Course, also students who are conducting experimental research as a part of pregraduation project. They also expressed similar objections related item analysis.

Previously, best item test quality is described based on each aspects analysis separately, are:

- a. Statistically, valid item defined as item with correlation coefficient (r value) in category very high and highly valid or r value is higher than r table.
- b. Statistically, reliable instrument is set of item test with correlation coefficient (r value) in category very high and highly reliable or r value is higher than r table.
- c. Difficulty index of item is stated in proportion value 0.00-1.00 (Aiken, 1994, p:66). According to Anwar (2009, p:61), for formative assessment, good item has index p 0.5. Tuckman D S (1992) in Syafril Anwar (2009, p:621) suggest that good item has p value from 0,33 - 0,67.
- d. Discriminancy power (D) index range are from -1,00 from +1,00. The higher of this index, the item is being better. Items with D value and further action subsequently are $D=0,4$ is ready used, $D=0,2-0,39$ need revision, $D < 0,20$ are disposed (Anwar, 2009, p:62).

Rarely, analysed items complete all of best for each criteria above. Teachers being confused to be guided by these criteria because there was no some certain rule or pattern that shows relationships among them. Crocker & Algina (1986) in Sumarna (2005), stated that reliability coefficient is influenced by several factor, one of them is difficulty index. Alen and Yen (1979) suggest that difficulty index has more dominant role among others. More difficult item determine more reliable, consistently. Arikunto (2005) described relationship between difficulty index and discriminancy power, that, item which difficulty index value 0.5 has possibility to get the highest value of discriminancy power. Separated study conducted by Fadilah, Alberida, Rahmawati (2011) informed that both difficulty index and discriminancy power do not have relationship each other.

Based on this condition, a study has been conducted to explore correlation among these aspect analysis in determining a best quality item, especially relationship between discriminancy power and item validity. The main question is "Does discriminancy power has relationship to item validity?".

RESEARCH METHOD

This descriptive correlational research used 12 samples, are sets of multiple choice test with characteristics: 1) Have been tested empirically or tested to testee, 2) Arranged by teacher or teacher candidates(educational program student), 3) Based on Curriculum 2006 and Basic Competence for second class of senior high school. The sample chosen by using purposive sampling. Data consist of a number of discriminancy power values and validity coefficient of each item. Data analysed statistically using correlation technique.

RESULT AND DISCUSSION

A. Relationship between discriminancy power and validity

In order to determine if there were relationship between items which discriminancy power is good with valid item, parametric correlation analysis was completed by Correlation Pearson, for 416 items dispersed in 12 samples. Majority (83,3%), there was a significant correlation between discriminancy power index and coefficient of validity. Contribution of discriminancy power to validity of item described by r^2 . See table 1.

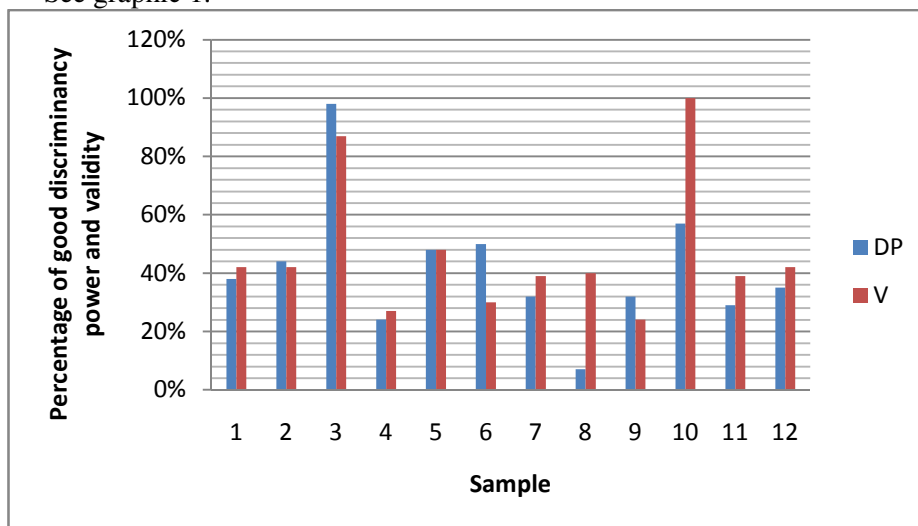
Table 1. Summary of correlation analysis result between discriminancy power index and item validity coefficient.

Sample	N	<i>r-coefficient</i> (α 0.01)	r^2 (%)
1 (SMS1)	44	.813**	66.1
2 (SMS2)	30	.909**	82.6
3 (SMS3)	36	.868**	73.9
4 (SMS4)	31	.614	43.0
5 (SME1)	20	.835**	69.7
6 (SME2)	20	.787**	61.9
7 (SME3)	20	.852**	72.6
8 (SME4)	20	.849**	72.1
9 (SGS1)	34	.740**	54.8
10 (SGS2)	38	.803**	64.5
11 (SGS3)	43	.779**	60.7
12 (SGS4)	50	.796**	63.4

**Correlation is significant at 0.01 level

B. Percentage of item with good discriminancy power and validity

Most highly relationship between discriminancy power and validity of items demonstrated that more items with good discriminancy power level would support the number of valid item. See graphic 1.



Relationship between discriminancy power and validity of items indicated by r -coefficient range from 0.740 to 0.909. Cohesiveness between discriminancy power and validity of items determined by r^2 , being approximate from 60.7% to 82.6%. It means that, item with good discriminancy power tend to have good validity level. Both discriminancy power and validity of items can be referred equally to judge items as ready-use item.

Previously, calculation of discriminancy power obtained from subtraction of upper students' answer proportions from lower students' answer proportion, result a proportion index which is categorized into some criteria. According to Arikunto (2008, p:218), the criteria is very good for 0.71 – 1.00, good for 0.41 – 0.70, sufficient for 0.21 – 0.40 and bad for 0.00 – 0.20.

Besides, it can be counted as explained by Mattlock-Hetzel (2011) and Wiesner (2011), "two indicators of the item's discrimination effectiveness are point biserial correlation and biserial correlation coefficient. The choice of correlation depends upon what kind of question we want to answer. The advantage of using discrimination coefficients over the discrimination index (D) is that every person taking the test is used to compute the discrimination coefficients and only 54% (27% upper + 27% lower) are used to compute the discrimination index, D. Point biserial. The point biserial (r_{pbis}) correlation is used to find out if the right people are getting the items right, and how much predictive power the item has and how it would contribute to predictions. Henrysson (1971) suggests that the r_{pbis} tells more about the predictive validity of the total test than does the biserial r , in that it tends to favor items of average difficulty. It is further suggested that the r_{pbis} is a combined measure of item-criterion relationship and of difficulty level. Biserial correlation. Biserial correlation coefficients (r_{bis}) are computed to determine whether the attribute or attributes measured by the criterion are also measured by the item and the extent to which the item measures them. The r_{bis} gives an estimate of the well-known Pearson product-moment correlation between the criterion score and the hypothesized item continuum when the item is dichotomized into right and wrong (Henrysson, 1971). Ebel and Frisbie (1986) state that the r_{bis} simply describes the relationship between scores on a test item (e.g., "0" or "1") and scores (e.g., "0", "1", ... "50") on the total test for all examinees. "The correlation of an item with the total test score (internal method) or with an external criterion

(external method) is yet another way to investigate the degree of item discrimination. There are a variety of correlational indexes depending on the nature of the variables. The most common correlation is the point biserial (rpb) correlation which is used when the criterion measure (e.g., total score) is continuous and the item scores are dichotomous (e.g., correct-incorrect). The point biserial correlation coefficient ranges from -1.00 to +1.00.

With both the index of discrimination and correlational indexes, a positive value indicates positive item discrimination, a negative value indicates negative discrimination, and low values indicate low or no discrimination. Items demonstrating high positive discrimination (e.g., over .50) would probably be retained while those with negative or low absolute values should be rejected, unless the item is checking for mastery and all or nearly all examinees are expected to mark the item correct. Items with a discrimination index of between .20 and .50 should probably be modified. Of course, overall item difficulty serves as a mediator of the index of discrimination; that is, easier items ordinarily have lower discrimination indexes because item difficulty truncates or suppresses the range of scores on one of the variables, thus lowering the resulting correlation or index.

Empirical validity oriented on using of a criteria as reference. Also, it is known as item validity. Principle of item validity analysis is any correlation between two variable, is correlation between each item score (1, 2, 3,, n) with total score. Most common correlational formula is *Product Moment* by Karl Pearson.

Based on information above, relationship between discriminancy power and validity of items caused by determination of discriminancy power and validity of items use the same technique, is applying correlational principle. A correlation could inform the nature of or form of its relationship. r coefficient obtained is positive, means that relationship is parallel. Increasing of discriminancy power cause increasing of item validity consistently. Wiesner (2011) told that item discrimination indexes are sometimes used as indexes of item validity. For nearly all tests, the most essential quality for items to have is the power of discrimination.

Intensity of relationship between discriminancy power and validity of items attained 82.6%. It can be interpreted that contribution of discriminancy power value reach 82.8 % in determining a valid item. Both item aspect can be referred to judge an item status, ready-use, revised, or disposed.

CONCLUSION AND SUGGESTION

Discriminancy power has relationship toward item validity significantly at 0.01%. The correlation is positive. Relationships intensity attained 82.6%. It summarised that item with better discriminancy power also has a better validity value. Both criteria can referred as basic consideration in judge an item status, ready-used, revised, or disposed.

The weakness of this study is item quality only determined quantatively, without consider hidden factors, such students' academic ability and psychological condition at testing time. Besides, it is suggested to explore this relationship by collecting larger sample and well controlled. Through this finding, researcher also suggest to use discriminancy power and validity criteria as main consideration to pick item test that will be tested to testee.

REFERENCES

- Aiken, Lewis R. 1994. *Psychological Testing and Assessment*, (Eight Edition), Boston: Allyn and Bacon.
- Anastasi. Anne and Urbina, Susana. 1997. *Psychological Testing*, (Seventh Edition). New

- Jersey: Prentice-Hall, Inc.
- Anwar, Syafri. 2009. *Penilaian Berbasis Kompetensi*. Padang: UNP Press.
- Arikunto, Suharsimi. 2005. *Dasar Dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.
- Darsono (2002). Belajar dan pembelajaran. Semarang. IKIP Semarang Press.
- Imran Zafar. 2008. *Item Analysis Assumptions, Measure of exam internal consistency (reliability) Kuder-Richardson 20 (KR20)*. *Artikel Juni 2008. Database Administrator, Assessment Unit, Dept of Medical Education. Ext. 47142*
- Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart and Winston, Inc.
- Fadilah, M, Alberida, H and Rahmawati D. 2011. Analisis Kriteria Keterpakaian Butir Soal Berdasarkan Hubungan antara Validitas, Reliabilitas, Tingkat Kesukaran, dan Daya Pembeda. *Research Report*. Biology Departement. State University Padang.
- Nitko, Anthony J. (1996). *Educational Assessment of Students, Second Edition*. Ohio: Merrill and Imprint of Prentice Hall Englewood Cliffs.
- Nunnally, Jum C. (1978). *Psychometric Theory*, Second Edition. New Delhi: Tata McGrawHill Publishing Company Limited.
- Nurkancana, Wayan dkk. 1983. *Evaluasi Pendidikan*. Surabaya: Usaha Nasional.
- Popham, James W. (1995). *Classroom Assessment: What Teachers Need to Know*, Boston: Allyn and Bacon.
- Purwanto, Ngalm. 1988. *Prinsip-prinsip dan Teknik Evaluasi Pengajaran*. Bandung: PT. Remaja Rosdakarya.
- Slameto. 1988. *Evaluasi Pendidikan*. Jakarta: Bina Aksara.
- Sudijono, Anas. 2008. *Evaluasi Pendidikan*. Jakarta: PT. Raja Grafindo Persada.
- Sudjana, Nana. 2008. *Penilaian Hasil Proses Belajar Mengajar*. Bandung: PT. Remaja Rosdakarya.
- Surapranata, Sumarna. 2005. *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Tes*. Bandung: PT Remaja Rosda Karya.
- Suryabrata, Sumadi. 1987. *Pengembangan Tes Hasil Belajar*. Jakarta: Rajawali.
- Syamsurizal. 2006. *Assessmen Pembelajaran*. Padang: PHK A2 Jurusan Biologi FMIPA UNP.
- Thorndike, Robert M. (1997). *Measurement and Evaluation in Psychology and Education*, Sixth Edition. Ohio: Merrill, an imprint of Prentice Hall.
- W. A. Mehrens dan I. J. Lehmann. 1973. *Measurement and Evaluation in Education and Psychology*. New York: Holt, Rinehart and Winston,
- Wiji Hariyanti. 2006. Analisis soal-soal Biologi Kleas VII SMP Semester gasal pada buku pegangan guru se-kabupaten pati. *Skripsi*. Tidak diterbitkan.FMIPA Unnes.
- <http://hamiddarmadi.blogspot.com/2011/05/validitas-tes-dan-item-tes-hasil.html>, hamid darmadi, validitas tes dan item tes hasil belajar
- repository.upi.edu/operator/upload/s_mat_0706564_chapter3.pdf
- www.uiowa.edu, 6 November 2011, Reading Your Test Analysis Report, Technical Bulletin #17
- <http://www.seattle-therapy.org/>, diakses 11 November 2011, Victor V. Wiesner, Ph.D., LPC-S, NCC, CCMHC, Item Analysis,)
- <http://www.google.com/>, diakses November 2011, Basic Concepts in Item and Test Analysis. Susan Matlock-Hetzel. Texas A&M University, January 1997